Predicting Protein Function with the Relative Backbone Position Kernel

Leander Schietgat, Thomas Fannes, Jan Ramon

Machine Learning Group, DTAI, K.U.Leuven
Contact: {leander.schietgat,thomas.fannes,jan.ramon}@cs.kuleuven.be

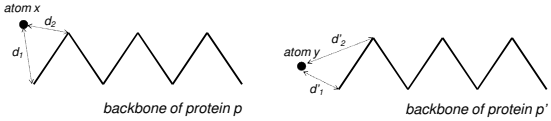
Machine Learning for Gene Function Prediction

- Proteins are important macromolecules
  - they play crucial roles in many biological processes
  - understanding their behaviour can lead to insights in disease development

- Machine learning and data mining
  - protein data is growing exponentially (mostly *primary or 1D* structures)
  - automatic classification of protein function is still an important challenge in bioinformatics
- Goal of this work
  - incorporate 3D protein data in machine learning methods
  - 3D data may contain useful information for the prediction task
  - challenges
    - 3D data may be hard to acquire
    - machine learning methods need upgrade to the 3D level

The Relative Backbone Position Kernel (RBPK)

- Kernel functions
  - measure similarities between objects
  - can be used in Support Vector Machines (SVMs) for classification, regression, …
  - advantages: accurate predictions, efficiently computable, …
- RBPK is a new protein kernel using 3D information
  - it discriminates between proteins by comparing Euclidean distances in 3D space between residue and backbone atoms
  - motivation: spatial features might be important for interactions with ligands or other proteins and will influence protein function



$$K_{3D}(p,p') = \sum_{x \in A(p) \setminus B(p)} \sum_{y \in \xi(x,p') \setminus B(p')} \lambda(x,y).k(x,y)$$

p and p' are 3D structures of proteins
A(p) is the set of atoms in protein p
B(p) is the set of atoms in the backbone of protein p

$$\xi(x,p) = \{y | y \in A(p), |b_1(x) - b_1(y)| < r\}$$

$b_1(x)$ is the Euclidean distance of atom x from backbone atom 1

$$\lambda(x,y) = \exp(-d(x,y)^2/\sigma^2)$$

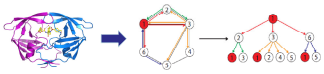
$$d(x,y) = \sqrt{\sum_{i=1}^{n}(b_i(x) - b_i(y))^2}$$

$b_i(x)$ is the Euclidean distance of atom x from backbone atom i
n is number of backbone atoms
k(x,y) = 1 if x = y and 0 otherwise

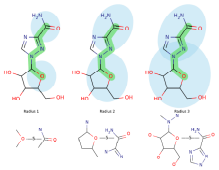
- Can be computationally demanding
  - controllable by parameter *r*

Related Work: Protein Kernels

- Fast Subtree Kernel (FSTK) [Shervashidze & Borgwardt, NIPS 2009]
  - proteins are converted into graphs
    - vertices are amino acids
    - edges between amino acids if close to each other



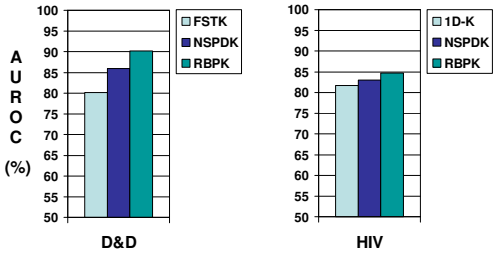
  - compares tree patterns extracted from neighbourhood of vertices
- Fast Neighbourhood Subgraph Pairwise Distance Kernel (NSPDK) [Costa & De Grave, ICML 2010]
  - decomposition kernel
  - compares extended pairwise neighbourhood graphs
    - several radii



None of these kernels use 3D information explicitly

Experimental Evaluation

- Experimental comparison of the protein kernels by using them in SVMs
  - 2 datasets
    - D&D: classification of enzymes
    - HIV: classification of resistance against IDV inhibitor
      - 3D structures are modelled on the sequences
      - use sequence kernel (1D-K) as baseline model
  - evaluation through AUROC reported under 10-fold cross-validation
- Results



Conclusions

- RBPK is a protein kernel that takes into account 3D information
  - by comparing distances between atoms in 3D space
- Preliminary experiments show that it reaches a state-of-the-art predictive performance
- Efficiency of RBPK needs to be improved
  - pruning strategies to limit the number of atom comparisons

MiGraNT project

K.U.Leuven
Department of Computer Science